Methods and reporting of systematic reviews of comparative accuracy need improvement: a methodological survey and proposed guidance

Yemisi Takwoingi[a,b], Christopher Partlett[c], Richard D. Riley[d], Chris Hyde[e], Jonathan J. Deeks[a,b]

[a]Institute of Applied Health Research, University of Birmingham, UK
[b]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK
[c]Nottingham Clinical Trials Unit, Faculty of Medicine and Health Science, University of Nottingham, Nottingham, UK
[d]Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK
[e]Exeter Test Group, College of Medicine and Health, University of Exeter, UK

Correspondence to:
Yemisi Takwoingi
Institute of Applied Health Research
University of Birmingham,
Birmingham, B15 2TT, UK
Email: y.takwoingi@bham.ac.uk
Telephone: 0121 4147833

Abstract

Objective: To examine methodological and reporting characteristics of systematic reviews and meta-analyses which compare diagnostic test accuracy (DTA) of multiple index tests, identify good practice, and develop guidance for better reporting.

Study design and setting: Methodological survey of 127 comparative or multiple tests reviews published in 74 different general medical and specialist journals. We summarised methods and reporting characteristics that are likely to differ between reviews of a single test and comparative reviews. We then developed guidance to enhance reporting of test comparisons in DTA reviews.

Results: Of 127 reviews, 16 (13%) reviews restricted study selection and test comparisons to comparative accuracy studies while the remaining 111 (87%) reviews included any study type. Fifty three reviews (42%) statistically compared test accuracy with only 18 (34%) of these using recommended methods. Reporting of several items—in particular the role of the index tests, test comparison strategy and limitations of indirect comparisons (i.e. comparisons involving any study type)—was deficient in many reviews. Five reviews with exemplary methods and reporting were identified.

Conclusions: Reporting quality of reviews which evaluate and compare multiple tests is poor. The guidance developed, complemented with the exemplars, can assist review authors in producing better quality comparative reviews.

Keywords: comparative accuracy; diagnostic accuracy; test accuracy; meta-analysis; systematic review; test comparison

Running title: Methods and reporting of systematic reviews of comparative accuracy

Word count (abstract): 200
Word count (main text only): 3322 (excluding tables, figures, captions and footnotes)

What is new?

Key findings

- Methods known to have methodological flaws are frequently used in reviews which evaluate and compare the accuracy of multiple tests. Reporting quality is variable but often poor.
- Test comparisons based on studies that have not directly compared the index tests are common in reviews but review authors fail to appreciate the potential for bias due to confounding.

What this adds to what was known?

- Guidance developed to promote better conduct and reporting of test comparisons in diagnostic accuracy reviews and to facilitate their appraisal. Exemplars also provided to assist review authors.

What is the implication and what should change now?

- To avoid misleading conclusions and recommendations, the methodological rigour and reporting of comparative reviews should be improved.
- Researchers and funders should recognise the merit of designing studies for obtaining reliable evidence about the relative accuracy of competing diagnostic tests.

1. Introduction

Medical tests are essential in guiding patient management decisions. Ideally, tests should only be recommended for routine clinical use based on evidence of their clinical performance (diagnostic accuracy) and clinical impact (benefits and harms) derived from relevant, high quality primary studies and systematic reviews. Systematic reviews and meta-analyses of diagnostic test accuracy (DTA) generally assess the performance of one index test at a time, thus providing a limited view of the test options available for a given condition and no information about the performance of alternatives. However, comparative reviews which compare the accuracy of two or more index tests are potentially more useful to clinicians and policy makers for guiding decision making about optimal test selection.

Since test evaluation is often limited to the assessment of test accuracy with limited or no regulatory requirement to demonstrate clinical impact,[1,2] it is vital that in the rapidly expanding evidence base, comparative accuracy reviews are conducted appropriately and well reported to avoid misleading conclusions and recommendations. Several reporting checklists have been developed to improve the transparency and reproducibility of medical research, including the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist [3] and PRISMA-DTA, the extension for DTA reviews.[4] Comparative accuracy reviews and meta-analyses are more challenging to perform than those of a single test; high quality reporting will enable assessment of the credibility of analysis methods and findings. Therefore, our aim was to summarise the methodological and reporting characteristics of comparative accuracy reviews, provide examples of good practice, and develop guidance for improving the reporting of test comparisons in future DTA reviews.

2. Methods

2.1. Terminology

To avoid confusion due to lack of standard terminology for types of test accuracy studies and systematic reviews, we describe here our choice of terminology. In Appendix Box 1 we provide a summary and other relevant definitions.

Unlike randomized controlled trials (RCTs) of interventions, which have a control arm, most test accuracy studies do not compare the index test with alternative index tests.[5] We used the term '*non-comparative*' to describe a primary study that evaluated a single index test or only one of the index tests being evaluated in a review, and '*comparative*' to describe a study that made a head-to-head comparison by comparing the accuracy of at least two index tests in the same study

population. A comparative study may either randomize patients to receive only one of the index tests (randomized design), or apply all the index tests to each patient (paired or within-subject design).[5] With both designs patients also receive the reference standard. For brevity, we will often refer to the index test simply as test.

We defined a comparative accuracy review as a review that met at least one of the following four criteria: (1) clear objective to compare the accuracy of at least two tests; (2) selected only comparative studies; (3) performed statistical analyses comparing the accuracy of all or a pair of tests; or (4) performed a direct (head-to-head) comparison of two tests. Reviews that assessed multiple tests but did not meet any of the four criteria were termed a multiple test review. Such reviews assess each test individually without making formal comparisons between tests and often include a large number of tests such as signs and symptoms from clinical examination. We included this category of reviews in order to be comprehensive and to avoid excluding reviews in the absence of established terminology.

The two main approaches for test comparisons in a DTA review are direct and indirect (between-study uncontrolled) comparisons (Appendix Figure 1). In a direct comparison only studies that have evaluated all the index tests are included in the comparison while an indirect comparison includes all eligible studies that have evaluated at least one of the index tests.

2.2. Data sources

We used an existing collection of 1023 systematic reviews published up to October 2012. The reviews were originally identified for an earlier empirical study using a previously described search strategy.[5] The reviews were identified by searching the Database of Abstracts of Reviews of Effects (DARE) for reviews with a structured abstract and the Cochrane Database of Systematic Reviews (CDSR issue 11, 2012). Reviews undergo quality appraisal before inclusion in DARE and so we expect reviews in DARE to be of higher quality than would be expected in the wider literature. We did not update the search because DARE is no longer being updated and we judged it unlikely that more recent reviews from the general literature would be of better methodological quality given the findings of recent empiric studies of DTA reviews.[6,7] Early publications (1980s and 1990s) of DTA reviews followed methodology for intervention reviews and key advances in methodology for DTA reviews were published between 1993 and 2005.[8] For these reasons, and to make allowance for dissemination of methods, reviews for the current study were limited to a five-year period from January 2008 to October 2012.

2.3. Eligibility criteria

All test accuracy reviews that evaluated at least two tests and included a meta-analysis were eligible. We excluded reviews where full text papers were unavailable, had insufficient data to determine study type (comparative or non-comparative), or where different tests were analysed together as a single test without separate meta-analysis results for each test.

2.4. Review selection and data extraction

Using a revised screening form from a previous empiric study, one assessor (YT or CP) assessed review eligibility by screening the abstract, followed by full text examination. When eligibility was unclear, the inclusion decision was made following discussion with a member of the author team (JD).

We scrutinized full text articles and their supplementary files. Data extraction was undertaken by one assessor (YT). To verify the data, a random subset of half of the included reviews was generated using the SURVEYSELECT procedure in SAS software, version 9.2 (SAS Institute, Cary, North Carolina). Data were extracted from these reviews by a second assessor. Any disagreements were discussed by the two assessors and agreement was achieved without having to involve a third person. We focused on methodological and reporting characteristics likely to differ between reviews of a single test and comparative reviews. We extracted data on general, methodological and reporting characteristics. These included data on target condition, tests evaluated, study design, and the analytical methods used for comparing tests and investigating differences between studies.

2.5. Development of test comparison reporting guidance

To identify a set of criteria, we used the list of methodological and reporting characteristics that we devised and the PRISMA-DTA checklist, combined with theoretical reasoning based on published methodological recommendations [8-10] and the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.[11] The criteria were selected to emphasise their importance for test comparisons when completing the PRISMA-DTA checklist for a comparative review.

2.6 Data analysis

We computed descriptive statistics for categorical variables as frequencies and percentages. Continuous variables were summarised using the median, range and interquartile range. Using the criteria and definition specified in section 2.1, we categorised reviews into comparative and multiple tests reviews. We subdivided comparative reviews into comparative reviews with and without a

statistical comparison because one of the key aspects that we examined was synthesis methods. Thus we summarised and presented our findings within three review categories. All data analyses were done using Stata SE version 13.0 (Stata-Corp, College Station, Texas, USA).

3. Results

The flow of reviews through the screening and selection process is shown in Figure 1. Of the 1023 reviews in the collection, 127 reviews met the inclusion criteria.

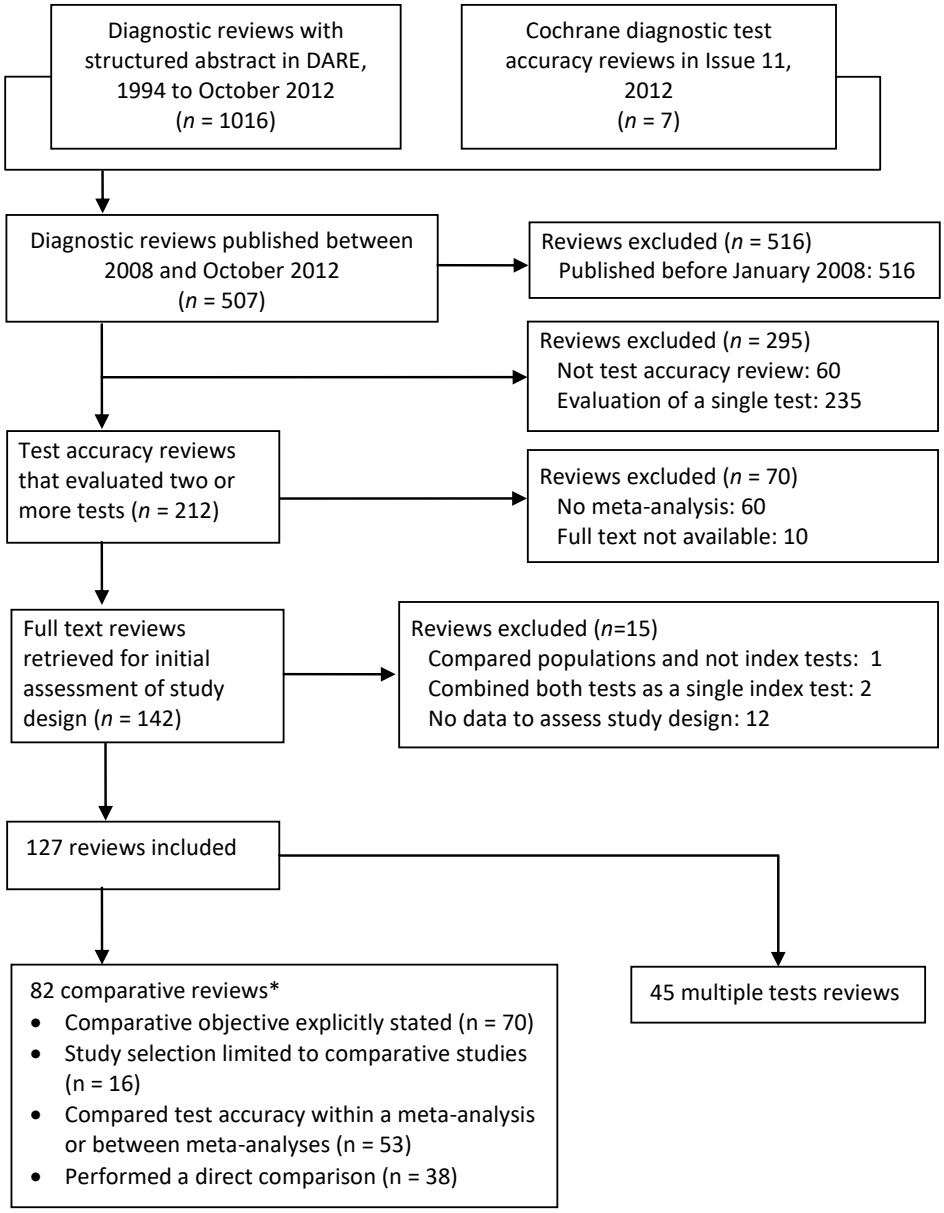


Figure 1. Flow of reviews through the selection process

*The 82 comparative accuracy reviews met at least one of the following four criteria: (1) clear objective to compare the accuracy of at least two tests; (2) selected only comparative studies; (3) performed statistical analyses comparing the accuracy of all or at least a pair of tests; or (4) performed a direct (head-to-head) comparison of two tests.

3.1. General characteristics

There were 82 comparative reviews and 45 multiple test reviews. Of the 82 comparative reviews, 53 (66%) formally compared test accuracy. Characteristics of the 127 reviews are summarised in Table 1. The reviews were published in 74 different journals, with the majority [93 {73%)] in specialist medical journals. The reviews covered a broad array of target conditions and test types, with neoplasms (37%), and imaging tests (43%) being the most frequently assessed target condition and test type. The median (interquartile range) number of comparative and non-comparative studies included per review were 6 (3 to 11) and 14 (3 to 24), respectively.

Table 1. Descriptive characteristics of 127 reviews of comparative accuracy and multiple tests

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical test performed to compare accuracy			
	Yes	No or unclear*		
Number of reviews	53 (42)	29 (23)	45 (35)	127
Year of publication				
2008	14 (26)	11 (38)	13 (29)	38 (30)
2009	6 (11)	10 (34)	8 (18)	24 (19)
2010	16 (30)	4 (14)	11 (24)	31 (24)
2011	13 (25)	3 (10)	7 (16)	23 (18)
2012†	4 (8)	1 (3)	6 (13)	11 (9)
Type of publication				
Cochrane review	3 (6)	1 (3)	1 (2)	5 (4)
General medical journal	5 (9)	5 (17)	13 (29)	23 (18)
Specialist medical journal	42 (79)	22 (76)	30 (64)	93 (73)
Technology assessment report	3 (6)	1 (3)	2 (4)	6 (5)
Number of tests evaluated				
2	20 (38)	14 (48)	12 (27)	46 (36)
3	12 (23)	6 (21)	4 (9)	22 (17)
4	8 (15)	3 (10)	4 (9)	15 (12)
≥5	13 (25)	6 (21)	25 (56)	44 (35)
Clinical topic (according to ICD-11 Version: 2018)				
Circulatory system	9 (17)	5 (17)	5 (11)	19 (15)
Digestive system	3 (6)	1 (3)	8 (18)	12 (9)
Infectious and parasitic diseases	3 (6)	4 (14)	9 (20)	16 (13)
Injury, poisoning and certain other consequences of external causes	2 (4)	1 (3)	2 (4)	5 (4)
Mental, behavioural or neurodevelopmental disorders	2 (4)	1 (3)	3 (7)	6 (5)
Musculoskeletal system and connective tissue	1 (2)	1 (3)	4 (9)	6 (5)
Neoplasms	28 (53)	12 (41)	7 (16)	47 (37)
Other ICD-11 codes‡	5 (9)	4 (14)	7 (16)	16 (13)
Type of tests evaluated				
Biopsy	0	1 (3)	0	1 (1)
Clinical and physical examination	5 (9)	3 (10)	15 (33)	23 (18)
Device	1 (2)	0	0	1 (1)

Characteristic	Comparative reviews		Multiple test reviews	Total
	Statistical test performed to compare accuracy			
	Yes	No or unclear*		
Imaging	32 (60)	13 (45)	9 (20)	54 (43)
Laboratory	8 (15)	8 (28)	12 (27)	28 (22)
RDT or POCT	1 (2)	0	4 (9)	5 (4)
Self-administered questionnaire	1 (2)	1 (3)	0	2 (2)
Combinations of any of the above§	5 (9)	3 (10)	5 (11)	13 (10)
Clinical purpose of the tests				
Diagnostic	42 (79)	23 (79)	44 (98)	109 (86)
Monitoring	1 (2)	1 (3)	0	2 (2)
Prognostic/prediction	0	1 (3)	0	1 (1)
Response to treatment	1 (2)	0	0	1 (1)
Screening	3 (6)	4 (14)	1 (2)	8 (6)
Staging	6 (11)	0	0	6 (5)
Number of test accuracy studies in reviews				
Median (range)	25 (6–103)	17 (5–82)	19 (3–79)	20 (3–103)
Interquartile range	14–43	11–32	12–24	12–34
Number of comparative studies				
Median (range)	7 (0–59)	6 (0–32)	4 (0–52)	6 (0–59)
Interquartile range	4–14	1–11	2–10	3–11
Number of non-comparative studies				
Median (range)	17 (0–98)	6 (0–79)	10 (0–76)	14 (0–98)
Interquartile range	6–32	0–27	5–20	3–24

ICD-11 = International Classification of Diseases, Eleventh Revision; RDT = Rapid diagnostic test; POCT = Point of care test.

*In 3 reviews, it was unclear whether a statistical comparison of test accuracy was done.

†Includes only studies published up to October 2012.

‡Includes 8 ICD-11 codes that had fewer than 5 reviews across the 3 groups.

§Tests evaluated in a review were not of the same type.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

3.2. Statistical characteristics

3.2.1. Use of comparative studies and test comparison strategies

Sixteen (13%) reviews restricted study selection and test comparisons to comparative studies while the remaining 111 (87%) reviews included any study type (Table 2). In 22 reviews (17%), both direct and indirect comparisons were performed with the direct comparisons performed as secondary analyses using pairs of tests for which data were available. Direct comparisons were not performed in 49 (39%) reviews even though comparative studies were available in 40 of the reviews and qualitative or quantitative syntheses would have been possible.

3.2.2. Methods for comparative meta-analysis and informal comparisons

We classified methods used in the 53 comparative reviews that statistically compared test accuracy into three main groups: (i) naïve comparison (19/53, 36%) which refers to a comparison where a statistical test, e.g. a Z-test, was used to compare summary estimates from separate meta-analysis

of one test with summary estimates from the meta-analysis of another test; (ii) univariate pooling of differences in sensitivity and specificity, or pooling of differences in the diagnostic odds ratio (6/53, 11%); and (iii) meta-regression by adding test type as a covariate to a meta-analytic model (23/53, 44%). For the remaining 5 (9%) reviews, the method used was unclear. Relative measures were used to summarise differences in accuracy in 18 of the 53 (34%) reviews.

For the remaining 29 comparative reviews that did not formally compare tests (i.e. through statistical quantification of the difference in accuracy, either via a p-value or estimate of the difference), three (10%) determined the statistical significance of differences in test accuracy based on whether or not confidence intervals overlapped, nine (31%) narratively compared tests, 14 (48%) did not perform a comparison and three (10%) were unclear.

Table 2. Strategies and methods for test comparisons

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
Number of reviews*	53 (42)	29 (23)	45 (35)	127 (100)
Study type				
Comparative only	8 (15)	8 (28)	0	16 (13)
Any study type	45 (85)	21 (72)	45 (100)	111 (87)
Test comparison strategy				
Direct comparison only	8 (15)	8 (28)	0	16 (13)
Indirect comparison only – comparative studies available	26 (49)	10 (34)	4 (9)	40 (32)
Indirect comparison only – no comparative studies available	2 (4)	6 (21)	1 (2)	9 (7)
Both direct and indirect comparison	17 (32)	5 (17)	0	22 (17)
None	0	0	40 (89)	40 (32)
Method used for test comparison†				
Meta-regression – hierarchical model	18 (34)	0	0	18 (14)
Meta-regression – SROC regression	2 (4)	0	0	2 (2)
Meta-regression – ANCOVA	2 (4)	0	0	2 (2)
Meta-regression – logistic regression	1 (2)	0	0	1 (1)
Univariate pooling of difference in sensitivity and specificity or DORs	6 (11)	0	0	6 (5)
Naïve (comparison of pooled estimates from separate meta-analyses)		0	0	
Z-test	15 (28)	0	0	15 (12)
Paired t-test	1 (2)	0	0	1 (1)
Unpaired t-test	1 (2)	0	0	1 (1)
Chi-squared test	1 (2)	0	0	1 (1)
Comparison of Q* statistic and their SEs‡	1 (2)	0	0	1 (1)
Overlapping confidence intervals	0	3 (10)	0	3 (2)
Narrative	0	9 (31)	4 (9)	13 (10)
None	0	14 (48)	40 (89)	54 (43)
Unclear	5 (9)	3 (10)	1 (2)	9 (7)

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
Relative measures used to summarise differences in test accuracy	18 (34)	0	0	18 (14)
Multiple thresholds included	13 (25)	12 (41)	17 (38)	42 (33)
If multiple thresholds included, were they accounted for in the comparative meta-analysis (meta-analysis at each threshold or fitted appropriate model)				
Yes	6 (46)	0	0	6 (46)
No	4 (31)	0	0	4 (31)
Unclear	3 (23)	0	0	3 (23)

ANCOVA = analysis of covariance; DOR = diagnostic odds ratio; SE = standard error; SROC = summary receiver operating characteristic.

*Numbers in parentheses are row percentages.

†These methods either involve a comparative meta-analysis or follow-on from a meta-analysis of each test individually.

‡Moses et al proposed the Q* statistic as an alternative to the area under the curve.[12] Q* is the point on the SROC curve where sensitivity is equal to specificity, i.e. the intersection of the summary curve and the line of symmetry.

Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

3.2.3. Investigations of heterogeneity

Investigations of heterogeneity were performed for individual tests in 67 (53%) reviews, of which 24 (36%) used meta-regression, 35 (52%) used subgroup analyses, and 8 (12%) used both methods (Table3). Amongst the 53 comparative reviews with a statistical comparison, 33 (62%) investigated heterogeneity. Five (15%) of the 33 reviews assessed the effect of potential confounders on relative accuracy using subgroup analyses (four reviews) or Bayesian bivariate meta- regression (one review).

Table 3. Investigations of heterogeneity in comparative and multiple test reviews

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
Number of reviews*	53 (42)	29 (23)	45 (35)	127 (100)
Formal investigation performed				
Yes – meta-regression and subgroup analyses	5 (9)	1 (3)	2 (4)	8 (6)
Yes – meta-regression	15 (28)	5 (17)	4 (9)	24 (19)
Yes – subgroup analyses	13 (25)	8 (28)	14 (31)	35 (28)
No – limited data	8 (15)	2 (7)	1 (2)	11 (9)
No – only tested for heterogeneity	3 (6)	8 (28)	16 (36)	27 (21)
No – nothing reported	7 (13)	5 (17)	8 (18)	20 (16)
Unclear	2 (4)	0	0	2 (2)
If yes above, was effect on relative accuracy also investigated?				
Yes	5 (15)	0	0	5 (15)
No	21 (64)	0	0	21 (64)

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
Planned but no data	1 (3)	0	0	1 (3)
Unclear	6 (18)	0	0	6 (18)

*Numbers in parentheses are row percentages.
Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

3.3. Presentation and reporting

Thirteen reviews (10%) used a reporting guideline (Table 4). Five reviews used PRISMA; four used QUORUM (Quality of Reporting of Meta-analyses), the precursor to PRISMA; one used both QUORUM and PRISMA; one used both STARD (Standards for the Reporting of Diagnostic accuracy) and MOOSE (Meta-analysis of Observational Studies in Epidemiology); and the remaining two stated they followed recommendations of the Cochrane DTA Working Group.

3.3.1. Summary of reporting quality and exemplars

Based on recommendations in the Cochrane Handbook,[13] five comparative reviews[14-18] were judged exemplary in terms of clarity of objectives and reporting of test comparison methods. A brief summary of the reviews is given in Appendix Table 1. Figure 2 summarises results for 10 reporting characteristics (derived from Table 4) for each of the 127 reviews. The figure clearly shows that the reporting of several items—in particular the role of the index tests, test comparison strategy and limitations of indirect comparisons—was deficient in many reviews. Further details are provided in sections 3.3.2 to 3.3.6.

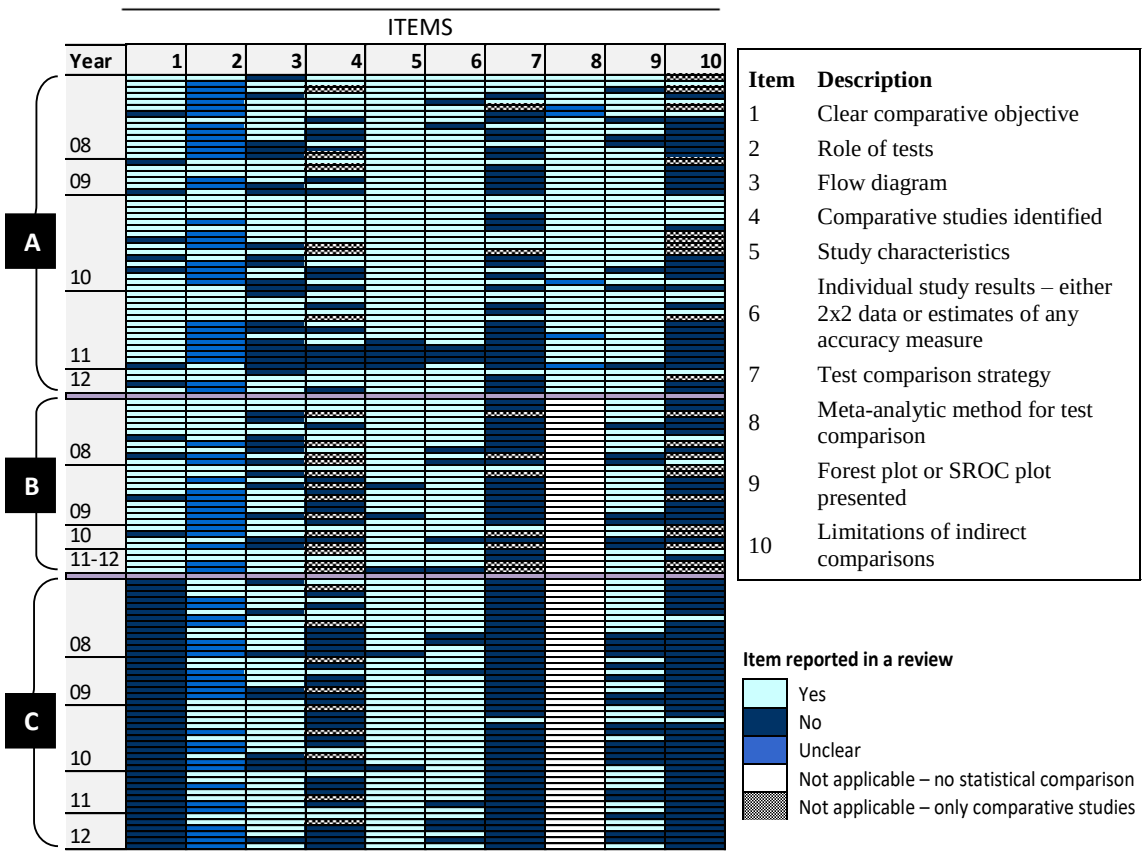


Figure 2: Reporting characteristics of 127 comparative and multiple test reviews

A– Comparative reviews with statistical analyses performed to compare accuracy; B – Comparative reviews without statistical analyses to compare accuracy; C – Multiple test reviews. The coloured cells in each row illustrate the reporting of the 10 items in each review. The box to the right of the figure gives the description of the reporting items. Reviews were ordered by year of publication and the number of missing items within each of the three review categories A to C. All multiple test reviews did not state a clear comparative objective (this was one of the four criteria used to classify the reviews as stated in section 2.1).

Table 4. Reporting and presentation characteristics of the reviews

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
Number of reviews*	53 (42)	29 (23)	45 (35)	127 (100)
Reporting guideline used	2 (4)	5 (17)	6 (13)	13 (10)
Clear comparative objective stated	45 (85)	25 (86)	0	70 (55)
Role of the tests				
Add-on	6 (11)	3 (10)	2 (4)	11 (9)
Replacement	8 (15)	6 (21)	6 (13)	20 (16)
Triage	4 (8)	1 (3)	11 (24)	16 (13)
Any two of the above	4 (8)	4 (14)	2 (4)	10 (8)
Unclear	31 (58)	15 (52)	24 (53)	70 (55)
Flow diagram presented				
Yes – included number of studies per test	11 (21)	6 (21)	8 (18)	25 (20)
Yes – excluded number of studies per test	21 (40)	12 (41)	28 (62)	61 (48)
No	21 (40)	11 (38)	9 (20)	41 (32)
Comparative studies identified				
Yes	31 (58)	9 (31)	9 (20)	49 (39)

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No or unclear		
No	16 (30)	7 (24)	27 (60)	50 (39)
No comparative studies in review	6 (11)	13 (45)	9 (20)	28 (22)
Study characteristics presented	48 (91)	26 (90)	43 (96)	117 (92)
Test comparison strategy				
Yes[†]	19 (36)	2 (7)	1 (2)	22 (17)
No[†]	32 (60)	20 (69)	44 (98)	96 (76)
No – included only comparative studies	2 (4)	7 (24)	0	9 (7)
Method used for test comparison[‡]				
Yes	48 (91)	NA	NA	48 (91)
Unclear	5 (9)	NA	NA	5 (9)
2x2 data for each study	30 (57)	10 (34)	14 (31)	54 (43)
Individual study estimates of test accuracy	46 (87)	25 (86)	36 (80)	107 (84)
Forest plot(s)	30 (57)	19 (66)	16 (36)	65 (51)
SROC plot				
SROC plot comparing summary points or curves for 2 or more tests	19 (36)	7 (26)	2 (4)	28 (22)
Separate SROC plot per test	17 (32)	11 (38)	19 (42)	47 (37)
No SROC plot	17 (32)	11 (38)	24 (53)	52 (41)
Limitations of indirect comparison acknowledged				
Yes	13 (25)	3 (10)	2 (4)	18 (14)
No	30 (57)	15 (52)	43 (96)	88 (69)
No but only comparative studies included	10 (19)	11 (38)	0	21 (17)

NA = not applicable; SROC = summary receiver characteristic operating.
[*]Numbers in parentheses are row percentages.
[†]These reviews included both comparative and non-comparative studies.
[‡]These methods either involve a comparative meta-analysis or follow-on from a meta-analysis of each test individually.
Numbers in parentheses are column percentages unless otherwise stated. Percentages may not add up to 100% because of rounding.

3.3.2. Review objectives and clinical pathway

A comparative objective was explicitly stated in 70 (55%) reviews (Table 4). It was possible to deduce the role of the tests in 57 (45%) reviews as add on, triage and/or replacement for an existing test. For 28 of the 57 (49%) reviews, the role was explicitly stated while we used implicit information in the background and discussion sections to make judgements for the remaining 29 (51%) reviews.

3.3.3. Study identification and characteristics

A flow diagram illustrating the selection of studies was not presented in 41 (32%) reviews (Table 4). In 61 (48%) reviews, a flow diagram was presented without the number of studies per test while 25 (20%) reviews presented a comprehensive flow diagram with the number of studies per test. Of these 25 reviews, the flow diagrams in five reviews[15,19-22] were notable examples. These flow diagrams clearly showed the number of studies included in the analysis of each test, and also indicated the number of comparative studies available. Of the 99 reviews that had at least one

comparative study, 50 (51%) reviews did not identify the comparative studies. Most of the reviews (92%) reported study characteristics though the detail reported varied.

3.3.4. Strategy for comparing test accuracy

Seventy three comparative reviews included both comparative and non-comparative studies and 21 (29%) of these reviews stated their strategy for comparing tests, i.e., direct and/or indirect comparisons (Table 4). Of the 21 reviews, 19 (90%) formally compared test accuracy.

3.3.5. Graphical presentation of test comparisons

A SROC plot showing results for two or more tests was presented in 28 (22%) reviews, 47 (37%) reviews showed each test on a separate SROC plot, and the remaining 52 (41%) reviews did not present a SROC plot (Table 4). Two multiple test reviews and seven comparative reviews without a formal test comparison presented a SROC plot showing a test comparison.

3.3.6. Limitations of indirect comparisons

Twenty one (17%) reviews restricted inclusion to comparative studies (Table 4). Of the remaining 106 reviews that included any study type, 18 (17%) acknowledged the limitations of indirect comparisons. Furthermore, 9 of these 18 reviews recommended that future primary studies should directly compare the performance of tests within the same patient population.

4. Discussion

4.1. Principal findings

The findings of our methodological survey showed considerable variation in methods and reporting. Despite the importance of clear review objectives, they were often poorly reported and the role of the tests was ambiguous in many reviews. Comparative studies ensure validity by comparing like-with-like thus avoiding confounding but only16 reviews (13%) restricted study selection to comparative studies. This may be due to scarcity of comparative studies.[5] It is worth noting that only two tests were evaluated in most (81%) of the 16 reviews that restricted inclusion to comparative studies.

The strategy adopted for test comparisons (direct comparisons and/or indirect comparisons) was not specified in many reviews. Further, the strategies that were specified varied considerably, reflecting a lack of understanding of the best methods for comparative accuracy meta-analysis. The validity of indirect comparisons largely depends on assumptions about study characteristics but

reviews did not always report study characteristics. To pool data for a direct or indirect comparison, the hierarchical methods recommended for comparative meta-analysis were not often used, with many reviews using methods known to have methodological flaws that can lead to invalid statistical inference. [13,23-25]

There are several potential sources of bias and variation in test accuracy studies,[26-28] and investigations of heterogeneity were commonly performed. However, the analyses were often done separately for each test rather than examining the effect jointly on all tests in a comparison. Understandably, the latter is rarely possible due to limited data. As empirical findings have shown that results of indirect comparisons are not always consistent with those or direct comparisons,[5] and adjusting for potential confounders in an indirect comparison will be uncommon, review findings should be carefully interpreted in the context of the quality and the strength of the evidence. Nevertheless, reviews seldom acknowledged the limitations of indirect comparisons.

4.2. Strengths and limitations

To our knowledge a comprehensive overview of reviews of comparative accuracy across different target conditions and types of tests has not been undertaken. We thoroughly examined a large sample of reviews published in a wide range of journals. Our classification of reviews was inclusive to enable a broad perspective of the literature and the generalisability of our findings. In addition to documenting review characteristics, we highlighted examples of good practice that review authors can use as exemplars. We also expanded relevant PRISMA-DTA items for reporting test comparisons in a DTA review.

Our study has limitations. First, the most recent review in our cohort of reviews was published in October 2012. Since the PRISMA-DTA checklist was published in January 2018, we did not update the collection as there had been no prior developments in reporting to suggest more recently published reviews would be better reported than older reviews. DARE is based on extensive searches of a wide array of databases and also includes grey literature. Given that for a review to be included in DARE it must meet certain quality criteria, the quality of the literature may be even poorer than we have shown. This view is supported by a study of 100 DTA reviews published between October 2017 and January 2018 which found that the reviews were not fully informative when assessed against the PRISMA-DTA and PRISMA-DTA for abstracts reporting guidelines.[7] Furthermore, we examined the use of six comparative meta-analysis methods that have been published since 2012 by checking their citations in Scopus.[29-34] Only one of the methods[32] had been

cited in a DTA review published in 2018. We also conducted a search of MEDLINE (Ovid) on July 31, 2019 to identify DTA reviews published in 2019 (Appendix 1). Of 151 records retrieved, 43 reviews met the inclusion criteria. The findings summarised in Appendix 1 show that test comparison methods and reporting remain suboptimal. Thus, our collection of reviews in this study reflects current practice.

Second, the assessment of the role of the tests was sometimes subjective and relied on the judgement of the assessor. Therefore, we only considered whether the item was reported or not, without assessing the quality of the description provided. We also discussed any uncertainty in a judgement before making a final decision.

4.3. Comparison with other studies

Previous research focused on systematic reviews of a single test or overview of any review type without detailed assessment of comparative reviews,[7,35,36] specific clinical area[37,38] or specific methodological issue.[39-41] Mallett et al[37] and Cruciani et al [38] concluded that conduct and reporting of DTA reviews in cancer and infectious diseases was poor. In an overview of DTA reviews published between 1987 and 2009, 36% of reviews that evaluated multiple tests reported statistical comparative analyses.[36] Similarly, 42% of our reviews reported such analyses.

4.4. Guidance and implications for research and practice

In Box 1 we provide reporting guidance for test comparisons to augment the PRISMA-DTA checklist and facilitate improvements in the reporting quality of comparative reviews. The guidance can also be used by peer reviewers and journal editors to appraise comparative DTA reviews. The challenges of a DTA review and the added complexity of test comparisons necessitate clear and complete reporting because of their increasing role in health technology assessment and clinical guideline development. Space constraints in journals are not an excuse for poor reporting because many journals publish online supplementary files. We noted that 56 (44%) reviews used supplementary files to provide additional data and information. Tutorial guides should be developed to assist review authors in navigating and understanding the complexity of DTA review methods. The Cochrane Screening and Diagnostic Tests Methods Group have already made contributions by providing freely available distance learning materials and tutorials on their website.

Since long-term RCTs of test-plus-treatment strategies which evaluate the benefits of a new test relative to current best practice are not always feasible[42,43] and are rare,[44] comparative accuracy

reviews are an important surrogate for guiding test selection and decision making. However, given the preponderance of indirect comparisons and paucity of comparative studies, there is a need to educate trialists, clinical investigators, funders and ethics committees about the merit of comparative studies for obtaining reliable evidence about the relative performance of competing diagnostic tests.

4.5. Conclusions

Comparative accuracy reviews can inform decisions about test selection but suboptimal conduct and reporting will compromise their validity and relevance. Complete and unambiguous reporting is therefore needed to enhance their use and minimise research waste. We advocate using the guidance we have provided as an adjunct to the PRISMA-DTA checklist to promote better conduct and reporting of test comparisons in DTA reviews.

Funding

Yemisi Takwoingi is funded by the UK National Institute for Health Research (NIHR) through a postdoctoral fellowship award (PDF-2017-10-059). Jonathan Deeks is a United Kingdom NIHR Senior Investigator Emeritus. Both Takwoingi and Deeks are supported by the NIHR Birmingham Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Declarations of interest: Takwoingi is an Associate Editor and Deeks is a Senior Editor of the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Deeks is the Editor-in-Chief, Takwoingi is a current member and Hyde is a former member of the Cochrane DTA Editorial Team. Takwoingi is a co-convenor of the Cochrane Screening and Diagnostic Tests Methods Group. Hyde, Deeks and Takwoingi are members of the PRISMA-DTA Advisory Group.

Box 1. Guidance for reporting test comparisons in systematic reviews of diagnostic accuracy

Item	Description (PRISMA-DTA items)*	Rationale and explanation
1	Role of tests in diagnostic pathway (3, D1)	Test evaluation requires a clear objective and definition of the intended use and role of a test within the context of a clinical pathway for a specific population with the target condition. The intended role of a test guides formulation of the review question and provides a framework for assessing test accuracy, including the choice of a comparator(s) and selection of studies. The role of a test is therefore important for understanding the context in which the tests will be used and the interpretation of the meta-analytic findings. The existing diagnostic pathway and the current or proposed role of the index test(s) in the pathway should be described. A new test may replace an existing one (replacement), be used before the existing test (triage) or after the existing test (add-on).[10]
2	Test comparison strategy (14)	Comparative studies are ideal but they are scarce.[5] An indirect between-study (uncontrolled) test comparison uses a different set of studies for each test and so does not ensure like-with-like comparisons; the difference in accuracy is prone to confounding due to differences in patient groups and study methods. Although direct comparisons based on only comparative studies are likely to ensure an unbiased comparison and enhance validity, such analyses may not always be feasible due to limited availability of comparative studies. Conversely, an indirect comparison uses all eligible studies that have evaluated at least one of the tests of interest thus maximising use of the available data (see Appendix Figure 1). If study selection is not limited to comparative studies and comparative studies are available, a direct comparison should be considered in addition to an indirect comparison. The direct comparison may be narrative or quantitative depending on the availability of comparative studies.
3	Meta-analytic methods (D2)	Hierarchical models which account for between-study correlation in sensitivity and specificity while also allowing for variability within and between studies are recommended for meta-analysis of test accuracy studies.[9,13] The two main hierarchical models are the bivariate and the hierarchical summary receiver characteristic operating (HSROC) models which focus on the estimation of summary points (summary sensitivities and specificities) and SROC curves respectively (see Appendix Figure 2).[45,46] For the summary point of a test to have a clinically meaningful interpretation, the analysis should be based on data at a given threshold. For the estimation of a SROC curve, data from all studies, regardless of threshold, can be included. As such test comparisons may be based on a comparison of summary points and/or SROC curves. For the estimation of a SROC curve using the HSROC model, one threshold per study is selected for inclusion in the analysis. If multiple cut-offs were considered, the description of methods should include how the cut-offs were selected and handled in the analyses. Methods have been proposed which allow inclusion of data from multiple thresholds for each study but the methods are yet to be applied to test comparisons.

Item	Description (PRISMA-DTA items)*	Rationale and explanation
4	Identification of included studies for each test (17)	Review complexity increases with increasing number of tests, target conditions, uses and/or target populations within a single review. Therefore, distinguishing between the different groups of studies that contribute to different analyses in the review enhances clarity. The PRISMA flow diagram can be extended to show the number of included studies for each test or group of tests if inclusion is not limited to comparative studies. The detail shown—individual tests or groups of tests, settings and populations—will depend on the volume of information and the ability of the review team to neatly summarise the information. If such a comprehensive flow diagram is not feasible, the studies contributing to the assessment of each test can be clearly identified in the manuscript in some other way. The source of the evidence should be declared by stating types of included studies. Studies contributing direct evidence should also be clearly identified in the review.
5	Study characteristics (18)	Relevant characteristics for each included study should be provided. This may be summarised in a table and should include elements of study design if eligibility was not restricted to specific design features. Heterogeneity is often observed in test accuracy reviews and differences between tests may be confounded by differences in study characteristics. Confounders can potentially be adjusted for in indirect test comparisons, though this is likely to be unachievable due to small number of studies and/or incomplete information on confounders. The effect of factors that may explain variation in test performance is typically assessed separately for each test.
6	Study estimates of test performance and graphical summaries e.g. forest plot and/or SROC plot (20)	It is desirable to report 2x2 data (number of true positives, false positives, false negatives and true negatives) and summary statistics of test performance from each included study. This may be done graphically (e.g. forest plots) or in tables. Such summaries of the data will inform the reader about the degree to which study specific estimates deviate from the overall summaries, as well as the size and precision of each study. It is plausible that study results for one test may be more consistent or precise than those of another test in an indirect comparison. In addition to forest plots, reviews may include SROC plots such as those shown in Appendix Figure 1 and Appendix Figure 2. A SROC plot of sensitivity against specificity displays the results of the included studies as points in ROC space. The plot can also show meta-analytic summaries such as SROC curves (panel B in Appendix Figure 2) or summary points (summary sensitivities and specificities) with corresponding confidence and/or prediction regions to illustrate uncertainty and heterogeneity, respectively (panel A in Appendix Figure 2). Ideally, results from a test comparison should be shown on a single SROC plot instead of showing the results for each test on a separate SROC plot. Furthermore, for pairwise direct comparisons, the pair of points representing the results of the two tests from each study can be identified on the plot by adding a connecting line between the points such as in panel B of Appendix Figure 1.
7	Limitations of the evidence	This is only applicable for reviews that include indirect comparisons. Be clear about the quality and strength of the evidence when interpreting the results, including limitations of including non-comparative studies in a test comparison. The results of indirect comparisons

Item	Description (PRISMA-DTA items)*	Rationale and explanation
	from indirect comparisons (24, 25)	should be carefully interpreted taking into account the possibility that differences in test performance may be confounded by clinical and/or methodological factors. This is essential because it is seldom feasible to assess the effect of potential confounders on relative accuracy.

*Related to the PRISMA-DTA item(s) indicated in parentheses.

References

1. Centre for Reviews and Dissemination. University of York. 2010. Accessed at http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp on 15 January 2019.
2. Horvath AR, Lord SJ, St John A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 2014;427:49-57.
3. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535.
4. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA 2018;319(4):388-396.
5. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013;158(7):544-5
6. Dehmoobad Sharifabadi A, Leeflang M, Treanor L, et al. Comparative reviews of diagnostic test accuracy in imaging research: evaluation of current practices. Eur Radiol. 2019.
7. Salameh JP, McInnes MDF, Moher D, et al. Completeness of Reporting of Systematic Reviews of Diagnostic Test Accuracy Based on the PRISMA-DTA Reporting Guideline. Clin Chem. 2019;65(2):291-301.
8. Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev 2013;2:82.
9. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149(12):889-897.
10. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332(7549):1089-1092.
11. Deeks JJ, Bossuyt PM, GGatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The Cochrane Collaboration. Accessed at https://methods.cochrane.org/sdt/handbook-dta-reviews on 25 January 2019.
12. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12(14):1293-1316.
13. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration; 2010. Accessed at https://methods.cochrane.org/sdt/handbook-dta-reviews on 25 January 2019.

14. Alldred SK, Deeks JJ, Guo B, Neilson JP, Alfirevic Z. Second trimester serum tests for Down's Syndrome screening. Cochrane Database of Syst Rev 2012;6:CD009925.
15. Wang LW, Fahim MA, Hayen A, et al. Cardiac testing for coronary artery disease in potential kidney transplant recipients. Cochrane Database of Syst Rev 2011(12):CD008691.
16. Pennant M, Takwoingi Y, Pennant L, et al. A systematic review of positron emission tomography (PET) and positron emission tomography/computed tomography (PET/CT) for the diagnosis of breast cancer recurrence. Health Technol Assess. 2010;14(50):1-103.
17. Williams GJ, Macaskill P, Chan SF, Turner RM, Hodson E, Craig JC. Absolute and relative accuracy of rapid urine tests for urinary tract infection in children: a meta-analysis. Lancet Infect Dis 2010;10(4):240-250.
18. Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. Ann Intern Med 2010;152(3):167-177.
19. Ewald B, Ewald D, Thakkinstian A, Attia J. Meta-analysis of B type natriuretic peptide and N-terminal pro B natriuretic peptide in the diagnosis of clinical heart failure and population screening for left ventricular systolic dysfunction. Intern Med J 2008;38(2):101-113.
20. Geersing GJ, Janssen KJ, Oudega R, et al. Excluding venous thromboembolism using point of care D-dimer tests in outpatients: a diagnostic meta-analysis. BMJ 2009;339:b2990.
21. Minion J, Leung E, Menzies D, Pai M. Microscopic-observation drug susceptibility and thin layer agar assays for the detection of drug resistant tuberculosis: a systematic review and meta-analysis. Lancet Infect Dis 2010;10(10):688-698.
22. Lucassen W, Geersing GJ, Erkens PM, et al. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. Ann Intern Med 2011;155(7):448-460.
23. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48(1):119-130; discussion 131-112.
24. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Medical decision making : an international journal of the Society for Med Decis Making 2008;28(5):621-638.
25. Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. Stat Methods Med Res 2016;25(4): 1596–1619.
26. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061-1066.
27. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-476.

28. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140(3):189-202.
29. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. Res Synth Methods 2014;5(4):294-312.
30. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. BMC Med Res Methodol 2015;15:70.
31. Nyaga VN, Aerts M, Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. Stat Methods Med Res 2018;27(6):1766-1784.
32. Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. Stat Methods Med Res 2018;27(5):1410-1421.
33. Ma X, Lian Q, Chu H, Ibrahim JG, Chen Y. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. Biostatistics 2018;19(1):87-102.
34. Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. J Clin Epidemiol 2018;99:64-74.
35. Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. BMC Med Res Methodol 2011;11:163.
36. Dahabreh IJ, Chung M, Kitsios GD, Terasawa T, Raman G, Tatsioni A, et al. Comprehensive overview of methods and reporting of meta-analyses of test accuracy. AHRQ publication no. 12-EHC044-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012.
37. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. BMJ 2006;333(7565):413.
38. Cruciani M, Mengoli C. An overview of meta-analyses of diagnostic tests in infectious diseases. Infect Dis Clin North Am. 2009;23(2):225-267.
39. Naaktgeboren CA, van Enst WA, Ochodo EA, et al. Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. J Clin Epidemiol 2014;67(11):1200-1209.
40. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005;9(12):1-113, iii.
41. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005;58(1):1-12.

42. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29(5):E1-E12.
43. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144(11):850-855.
44. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol 2012;65(3):282-287.
45. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-990.
46. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20(19):2865-2884.
47. Abba K, Deeks JJ, Olliaro P, et al. Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries. Cochrane Database of Syst Rev 2011(7):CD008122.
48. Carvalho AF, Takwoingi Y, Sales PM, et al. Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies. J Affect Disord 2014;172C:337-346.

Appendix 1. Supplementary methods and findings

Given the age of the study cohort of reviews, to ascertain the applicability of our conclusions and proposed guidance to current practice, we examined systematic reviews of test accuracy published in 2019. The aim of this post hoc assessment was to provide a brief overview of reviews published in 2019, focussing on test comparison methods and the reporting items highlighted in the guidance proposed in Box 1. A summary of the search and findings is given below.

Search strategy, study selection and data extraction

We performed a search of MEDLINE (Ovid) on 31st July 2019 to identify test accuracy reviews published between 1st January and 31st July 2019. Since PRISMA-DTA recommends that the title of a diagnostic test accuracy (DTA) systematic review should identify the report as a systematic review (meta-analysis) of DTA studies, we used the search strategy outlined below to obtain a snapshot of reviews published in 2019.

1 systematic review.m_titl.
2 meta-analysis.m_titl.
3 accuracy.m_titl.
4 1 OR 2
5 3 AND 4
6 limit 5 to yr="2019"

We applied the same eligibility criteria and selection process as described in the manuscript. One of three assessors (YT, CP and CH) screened titles, abstract and, if necessary, the full text to determine eligibility. Data extraction was performed by a single assessor (YT or CP). Any uncertainties were resolved by the assessors through discussion without the need to involve another member of the author team.

Search results

Of the 151 records identified by the search, 43 reviews assessed the accuracy of more than one index test (Figure A1). Based on our terminology for classifying the reviews, 37 reviews were judged to be comparative while the remaining 6 were multiple test reviews.

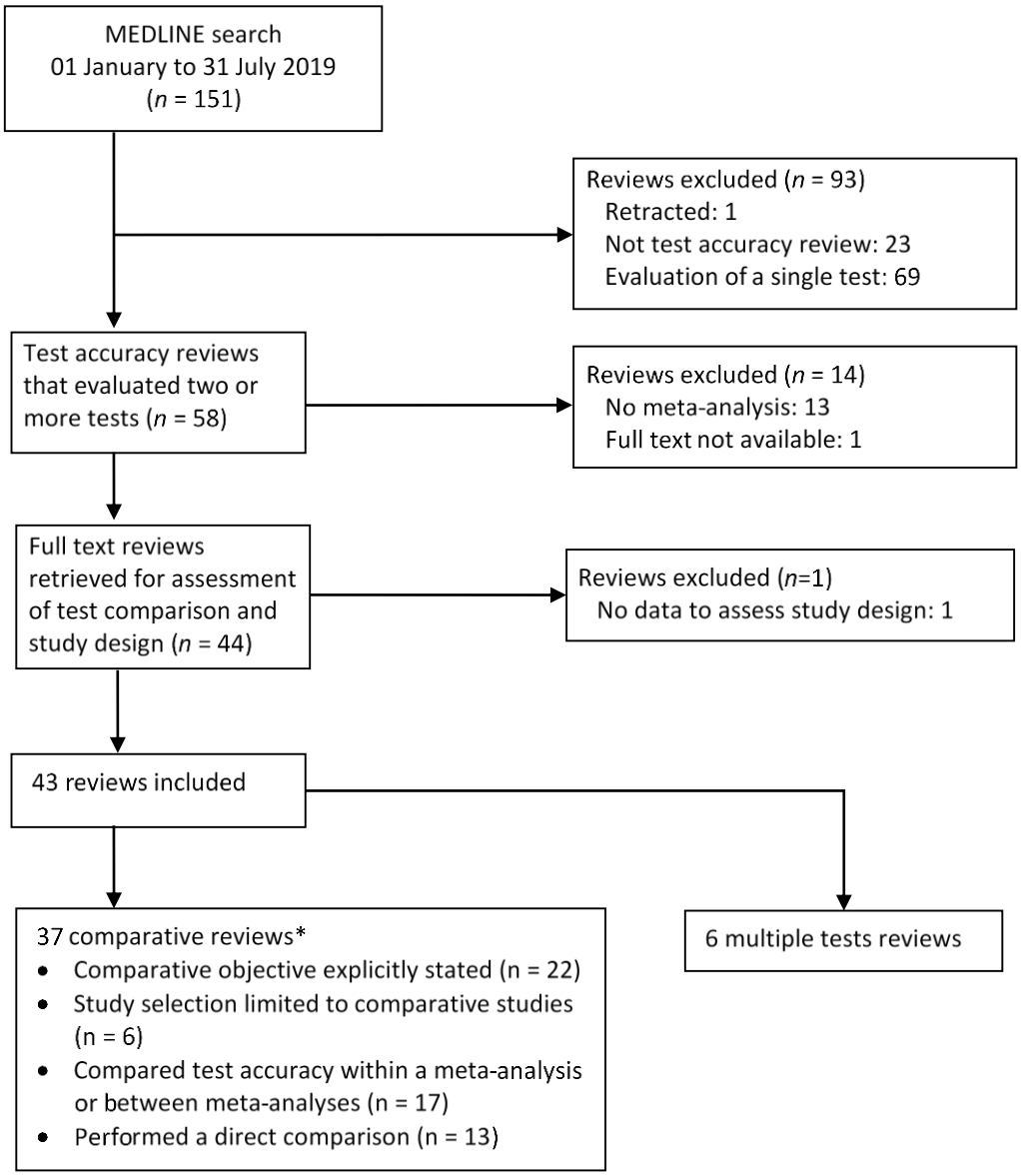


Figure A1. Flow of reviews through the selection process to identify eligible reviews published in 2019

*The 37 comparative accuracy reviews met at least one of the following four criteria: (1) clear objective to compare the accuracy of at least two tests; (2) selected only comparative studies; (3) performed statistical analyses comparing the accuracy of all or at least a pair of tests; or (4) performed a direct (head-to-head) comparison of two tests.

Findings

The 43 reviews were published in 39 different journals—33 specialist and six general medical journals. The number of tests evaluated in the reviews ranged between two and 17, with most reviews evaluating two (18/43; 42%) or three (7/43; 16%) tests.

Test comparison methods are summarised in Table A1 and reporting characteristics in Table A2. In this 2019 review cohort, 14% of the reviews restricted study selection to only comparative studies (Table A1), similar to the proportion (13%) for the DARE cohort of reviews (Table 2). We found that 10% of the 2019 cohort included only direct comparisons compared with 13% for the DARE cohort published between 2008 and 2012. The proportion of reviews using the recommended hierarchical meta-regression approach for test comparisons was the same (14%) for both review cohorts. Reporting of several items is still deficient in 2019 reviews (Table A2). For example, the test comparison strategy was not reported in 76% (Table 4) of the DARE cohort and 74% (Table A2) of the 2019 cohort. Reviews published in 2019 still mainly relied on indirect comparisons yet such reviews did not typically address the limitations of indirect comparisons; 69% of the DARE cohort and 57% of the 2019 cohort did not acknowledge limitations.

Table A1. Strategies and methods for test comparisons in a cohort of reviews published in 2019

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No		
Number of reviews*	20 (47)	17 (40)	6 (14)	43 (100)
Number of tests evaluated				
2	11 (55)	5 (29)	2 (33)	18 (42)
3	4 (20)	3 (18)	0	7 (16)
4	0	0	0	0
≥5	5 (25)	9 (53)	4 (67)	18 (42)
Study type				
Comparative only	4 (20)	2 (12)	0	6 (14)
Any study type	16 (80)	15 (88)	6 (100)	37 (86)
Test comparison strategy				
Direct comparison only	8 (40)	2 (12)	0	10 (24)
Indirect comparison only – comparative studies available	7 (35)	11 (65)	0	18 (42)
Indirect comparison only – no comparative studies available	2 (10)	2 (12)	0	4 (9)
Both direct and indirect comparison	2 (10)	1 (6)	0	3 (7)
Unclear	1 (5)	0	0	1 (2)
None	0	1 (6)	6 (100)	7 (16)
Method used for test comparison				
Meta-regression – hierarchical model	6 (30)	0	0	6 (14)
Network meta-analysis†	1 (5)	0	0	1 (2)
Univariate pooling of difference in sensitivity and specificity	1 (5)	0	0	1 (2)
Naïve (comparison of pooled estimates from	1 (5)	0	0	1 (2)

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No		
separate meta-analyses) using t-test				
Comparison of area under the curve	4 (20)	0	0	4 (9)
Overlapping confidence intervals	2 (10)	0	0	2 (5)
Overlapping prediction regions	1 (5)	0	0	1 (2)
Narrative	0	14 (82)	0	14 (33)
None	0	3 (18)	6 (100)	9 (21)
Unclear	4 (20)	0	0	4 (9)

NA = not applicable; SROC = summary receiver characteristic operating.

*Numbers in parentheses are row percentages.

†A review of 2 index tests stated that network meta-analysis was done.[1] However, details were not given and no method was cited. The network plots showed the 2 index tests and the reference standard. Based on the available information it is unclear if this was an appropriate network meta-analysis.

Table A2. Reporting and presentation characteristics of a cohort of reviews published in 2019

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No		
Number of reviews*	20 (47)	17 (40)	6 (14)	43 (100)
Reporting guideline used†	13 (65)	9 (53)	2 (33)	24 (56)
Clear comparative objective stated	14 (70)	8 (47)	0	22 (51)
Role of the tests				
Add-on	1 (5)	1 (6)	0	2 (5)
Replacement	3 (15)	2 (12)	0	5 (12)
Triage	1 (5)	3 (18)	1 (17)	5 (12)
Add on or replacement	1 (5)	0	0	1 (2)
Unclear	14 (70)	11 (65)	5 (83)	30 (70)
Flow diagram presented				
Yes – included number of studies per test	4 (20)	2 (12)	0	6 (14)
Yes – excluded number of studies per test	16 (80)	14 (82)	6 (100)	36 (84)
No	0	1 (6)	0	1 (2)
Comparative studies identified				
Yes	7 (35)	2 (12)	1 (17)	10 (23)
No	7 (35)	10 (59)	4 (67)	21 (49)
No comparative studies in review	6 (30)	5 (29)	1 (17)	12 (28)
Test comparison strategy				
Yes	9 (45)	2 (12)	0	11 (26)
No	11 (55)	15 (88)	6 (100)	32 (74)
Method used for test comparison				
Yes	15 (75)	NA	NA	15 (75)
Unclear	5 (25)	NA	NA	5 (25)
2x2 data for each study	9 (45)	8 (47)	3 (50)	20 (47)
Individual study estimates of test accuracy	17 (85)	14 (82)	5 (83)	36 (84)
Forest plot(s)	18 (90)	13 (76)	5 (83)	36 (84)
SROC plot				
SROC plot comparing summary points or curves for 2 or more tests	11 (55)	0	0	11 (26)
Separate SROC plot per test	7 (35)	9 (53)	3 (50)	19 (44)
SROC plot for one test only	1 (5)	3 (18)	1 (17)	5 (12)
No SROC plot	1 (5)	5 (29)	2 (33)	8 (19)
Limitations of indirect comparison acknowledged				

Characteristic	Comparative reviews Statistical analyses to compare test accuracy		Multiple test reviews	Total
	Yes	No		
Yes	2 (10)	2 (12)	0	4 (9)
No	10 (50)	13 (76)	6 (100)	29 (57)
No, only comparative studies included	8 (40)	2 (12)	0	10 (23)

NA = not applicable; SROC = summary receiver characteristic operating.

*Numbers in parentheses are row percentages.

†PRISMA-DTA and PRISMA reporting guidelines were used in six and 18 reviews respectively. The 19 remaining reviews did not state that a reporting guideline was used. One of the reviews that used PRISMA also used the MOOSE (Meta-analysis Of Observational Studies in Epidemiology) guideline.

References

1. Huang R, Jiang L, Xu Y, et al. Comparative Diagnostic Accuracy of Contrast-Enhanced Ultrasound and Shear Wave Elastography in Differentiating Benign and Malignant Lesions: A Network Meta-Analysis. *Front Oncol.* 2019;9:102.

Appendix Box 1. Definition of test accuracy terms

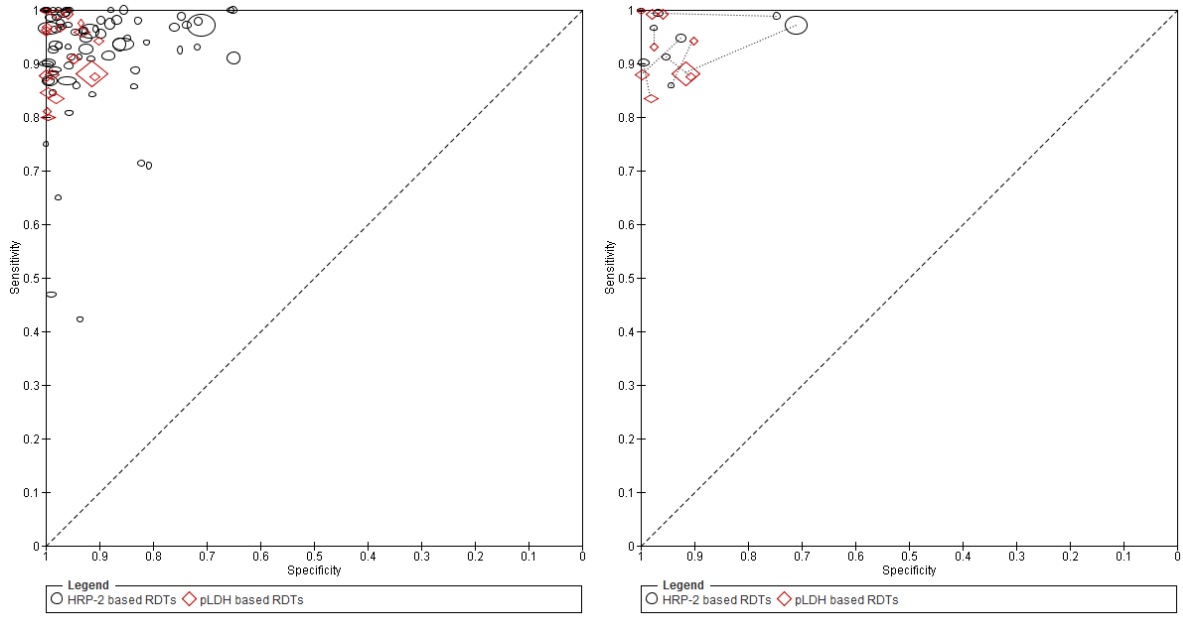
Term	Definition
Comparative accuracy study	A primary study that compared the accuracy of two or more index tests in the same study population by randomizing participants to only one of the index tests (randomized design), or by applying all the index tests to each participant (paired or within-subject design).
Comparative accuracy review	A diagnostic test accuracy review with a clear objective to compare the accuracy of at least two tests; limited study selection to only comparative studies; performed a direct (head-to-head) comparison of two tests; and/or performed a statistical comparison of test accuracy.
Diagnostic odds ratio (DOR)	Ratio of the odds of positivity in those who have the target condition compared to the odds of positivity in those without the condition.
Direct comparison	In a comparative diagnostic test accuracy review, a direct comparison includes only studies that have evaluated all the index tests being compared in the test comparison. Thus differences between studies in clinical or methodological characteristics are less likely to confound differences in accuracy.
Index test	The new or existing test of interest that is being evaluated.
Indirect comparison	In a comparative diagnostic test accuracy review, an indirect comparison includes all eligible studies that have evaluated at least one of the index tests in the test comparison. Thus indirect comparisons maximize use of the available data but are prone to confounding.
Multiple test review	A diagnostic test accuracy review that does not have an explicit objective to compare the accuracy of two or more but assessed the accuracy of each test individually without making any formal comparison between tests.
Non-comparative accuracy study	A primary study that evaluated the accuracy of a single index test or the accuracy of only one of the index tests being evaluated in a comparative accuracy review.
Q*	Point on the SROC curve where sensitivity is equal to specificity, i.e. the intersection of the summary curve and the line of symmetry (downward diagonal line running from the top left corner of the SROC plot to the bottom right corner).
Reference standard	The best way of verifying the presence or absence of the target condition. It may be a single test or a combination of tests and clinical information.
Sensitivity	Proportion of those with the target condition who have positive index test results.
Specificity	Proportion of those without the target condition who have negative index test results.
SROC plot	The summary receiver operating characteristic (SROC) plot is a scatterplot of sensitivity against specificity which displays the results of the studies included in an analysis from a diagnostic test accuracy review as points in receiver operating characteristic (ROC) space.

Target condition	The condition, clinical state or pathological disease that is to be detected or excluded.
Test accuracy	The ability of a test to discriminate between those who have and those who do not have the target condition. Test accuracy is estimated by comparing results of an index test with a reference standard, sometimes known as a 'gold' standard.

Appendix Figure 1. Comparison of test accuracy on SROC plots

A. Indirect comparison

B. Direct comparison



HRP-2= histidine-rich protein-2; pLDH= plasmodium lactate dehydrogenase; RDT = rapid diagnostic test.
For each test on a SROC plot, each symbol represents the pair of sensitivity and specificity from a study. To reflect the precision of sensitivity and specificity in each study, the size of the symbols was scaled by the sample sizes for those with and those without the target condition. Panel A shows an indirect comparison using data from the P. falciparum malaria review by Abba et al 2011.[47] The indirect comparison included 75 HRP-2 based RDT studies and 19 pLDH based RDT studies. Of these, nine studies compared both tests in the same patients (I.e. paired or within subject design) and were included in the direct comparison shown in panel B. Point estimates for an HRP-2 based RDT and a pLDH based RDT from the same study are connected by a dotted line.

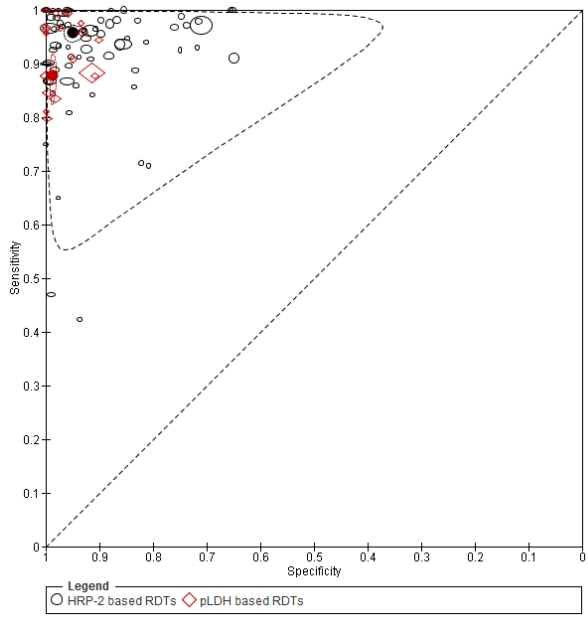
Appendix Table 1. Summary of methodological and reporting characteristics of five exemplar comparative reviews

Review	Publication type	Review objectives	Test comparison and meta-analysis	Comments
Alldred 2012[14]	Cochrane	To estimate and compare the accuracy of second trimester serum markers for the detection of Down's syndrome, both as individual markers and as combinations of markers.	Comprehensive description of statistical methods for both direct and indirect comparisons, including the strategy for handling multiple thresholds. HSROC meta-regression models were used to formally compare test accuracy. Relative accuracy was expressed in terms of the relative diagnostic odds ratio as appropriate.	A well-structured, large and complex review due to the number of tests, test combinations and thresholds included.
Wang 2011[15]	Cochrane	To assess the diagnostic accuracy of non-invasive cardiac screening tests versus coronary angiography in potential kidney transplant recipients. Diagnostic accuracy was compared between screening tests.	Clear and detailed description of test comparison strategy and methods. Both direct and indirect comparisons were planned. Test accuracy was statistically compared in a HSROC meta-regression model.	Included an exemplary flow diagram.
Pennant 2010[16]	Technology assessment report	To assess the incremental diagnostic accuracy of PET and PET/CT compared with existing diagnostic strategies and to compare the diagnostic accuracy of PET and PET/CT for the diagnosis of breast cancer recurrence	Rationale given for the test comparisons performed. For each pairwise comparison of imaging modalities, a bivariate meta-regression model was used to compare test accuracy. Relative accuracy reported using relative sensitivities and relative specificities.	Forest plots showing the pair of test accuracy estimates from each study were presented.
Williams 2010[17]	Specialist	To assess whether rapid urine tests were sufficiently sensitive to avoid urine culture in children with negative results and to compare the accuracy of dipsticks with microscopy.	Detailed test comparison strategy and methods. Differences in thresholds for test positivity were accounted for. Direct comparisons were performed using a HSROC model and meta-regression. Where appropriate, the relative diagnostic odds ratio was used to summarise relative accuracy.	A very detailed description of the methods.

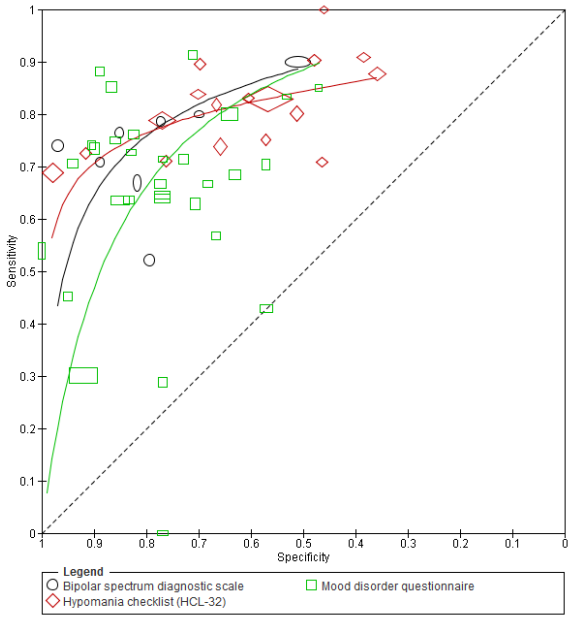
Review	Publication type	Review objectives	Test comparison and meta-analysis	Comments
Schuetz 2010[18]	General medical journal	To compare CT and MRI for ruling out clinically significant coronary artery disease (CAD) in adults with suspected or known CAD.	A bivariate meta-regression model was used to compare test accuracy. Both direct and indirect comparisons were done though the test comparison strategy was not specified in the methods.	Review findings interpreted with caution due to limited evidence from comparative studies.

Appendix Figure 2. Comparison of test accuracy using summary points or summary curves

A. Comparison of summary points



B. Comparison of summary curves



HRP-2= histidine-rich protein-2; pLDH= plasmodium lactate dehydrogenase; RDT = rapid diagnostic test.
For each test on a SROC plot, each symbol represents the pair of sensitivity and specificity from a study. The size of each symbol was scaled according to the precision of sensitivity and specificity in the study. Panel A shows a comparison of summary points using the P. falciparum malaria review by Abba et al 2011.[47] The indirect comparison included 75 HRP-2 based RDT studies and 19 pLDH based RDT studies. The solid circles (summary points) on the SROC plot represent the summary estimates of sensitivity and specificity for each test. Each summary point is surrounded by a dotted line representing the 95% confidence region and a dashed line representing the 95% prediction region (the region within which one is 95% certain the results of a new study will lie). Panel B shows a comparison of summary curves using the bipolar disorder review by Carvalho et al 2014.[48] The indirect comparison included 44 studies that evaluated the diagnostic accuracy of the mood disorder questionnaire (30 studies), the bipolar spectrum diagnostic scale (8 studies) and the hypomania checklist (HCL-32, 17 studies) for detection of any type of bipolar disorder in a mental setting. Each summary curve was drawn restricted to the range of specificities from included studies that evaluated the test.